複数の圧縮プログラムを用いた近代日本文学の著者推定

安　形　　輝

Authorship Attribution Using Data Compression Program for Modern Japanese Literature

Teru Agata

Abstract

The present study describes an experiment on authorship attribution using data from modern Japanese literature with a combination of CIR(compression improvement ratio) method and 16 compression program. The results of the experiment show (1) high performance compression programs give an average precision ratio of almost 100% regardless of algorithms, (2) with high performance compression programs, shortened data causes slight performance degradation, (3) an average precision ratio on authorship attribution highly correlates with the compression ratio of the program.

1. 圧縮プログラムを用いた著者推定

1. 1　著者推定

著者推定とは、作者不明のデータがあった場合にデータの特徴から著者を推定することであり、コンピュータの登場以前から様々な手法が提案され（村上 1994）、継続的に研究がなされてきた比較的活発な研究領域といえる。

時代的に古い資料のなかには、著作者が不明である文献や、作品群の著作者の同一性が問題となっている文献が少なからず存在している。前者を対象とした事例としては旧約聖書の著者推定があり（フリードマン 1989）、

後者を対象とした例としては日蓮が本当に著したのかが疑わしいとされている文献の真贋判定（村上 2002）が挙げられる。

著者推定や真贋判定は、特に文学研究において重要な研究領域の一つである。例えば、近代の著名作家の未公開作の発見時の真贋鑑定（細江 1988）といった事例がある。一方で、裁判における被告人の上申書と日記の作成者の同一性の検証（森 2001）といった著者推定の応用事例は、学術面からだけでなく実社会からの需要も高いことを示していると考えられる。近年ではデータの長さが極めて短いオンラインメッセージに関する著者推定についても検討されている（Zheng ら 2006）。

他の研究領域との関係からみると、著者推定は文体的特徴から類似データを識別するため、情報検索や自動分類と共通の枠組みを持つといえ、研究成果は互いに応用可能な場合が多い。例えば、佐藤ら（佐藤ほか 2002）はウェブ上の情報源間の自動分類に著名な著者判定手法である Tankard の手法（Tankard 1986）を応用している。

1．2　圧縮プログラムを用いた著者推定

圧縮プログラムあるいはアーカイバは、本来、データ中の冗長な部分を識別し、より短いデータに置き換えることによって全体のサイズを縮小し、外部記憶装置に占める容量を節約したり、あるいは、通信にかかる時間を短縮したりすることを目的としている。近年、その圧縮プログラムを本来の圧縮用途ではなく、類似データの識別に応用する試みが行われはじめている。

圧縮プログラムを応用した類似データの同定の基本的な考え方は、非常に単純なものである。二つのデータがあったときに、データ同士が類似していればしているほど、共通する冗長な部分が多くなる。そこで、ある二つのデータを連結した（二つのデータを単純に並置し、一つのファイルとした）ものを圧縮プログラムに投入したときに、共通する冗長な部分が多ければ、より高い圧縮率で圧縮ができると考えられる。つまり、多くのデータの中

で、圧縮プログラムがより小さいサイズまで圧縮できる二つのデータはより類似していることになる。いくつかのバリエーションはあるが、基本的にはこの原理を応用して圧縮サイズからデータ同士の類似度を測定する。

圧縮プログラムを類似度測定や自動分類に応用するアイディアは、圧縮アルゴリズムの考案者や圧縮プログラムの開発者には自明の手法であったという記述も存在する（奥村 2006）。しかし、近年になるまで実際に類似度測定に応用した研究事例として広く認知されたものはなかったため、以下の研究が読者数の多い学術誌や一般誌に掲載されるまでは、圧縮関係の研究者以外にはあまり知られていない手法であったといえる。

この手法に関する有名な研究として、Dario Benedetto らの“Language Trees and Zipping”（Benedetto ら 2002）がある。この文献中で、彼らはZIP 系列の圧縮プログラムによる自動分類や類似データの同定手法を提案し（以下、Benedetto らの手法）、DNA 配列の類似度測定、言語不明データの言語識別、著者不明データの著者推定に関する実験を行った結果を簡単に紹介した。著者推定に関しては 90 文献（Liber Liber. “The Manuzio plan”. <http://www.liberliber.it/>）から構成されるコーパスに対して著者推定実験を行い、93.3% という高い精度を得た。しかし、実験環境に関して詳細な記述がなく、著者推定に関する既往研究と同様の実験データを用いてもいない。そのため、実験結果の比較をすることができず、さらに実験の再現も難しい。

O.V. Kukushkina らは、Benedetto らに先んじて 2000 年に同様の手法で圧縮プログラムを応用したテキストの著者推定に関する実験を行った（Kukushkina ら 2000）。彼らの実験結果によれば、最も精度の高い圧縮プログラムは、マルコフ連鎖を応用した手法を超える高い精度を示した。しかし、この実験は、彼らが提案した手法の有効性を検証するためであり、Appendix-A として追加的かつ簡単に記述されただけである。そのため、本文にはマルコフ連鎖を応用した彼らの提案する手法とその実験結果しか書かれていない。ロシア語論文であったこと、Appendix であったことな

どから、*Physical Review Letters* 誌で Benedetto らの論文に関する議論において英訳ウェブページが紹介されるまでは、それほど認知されていなかった研究であると考えられる。

日本語データに関しては、内山和也（内山 2002）が Benedetto らの手法を用いて 7 人の書き手による日本語学術論文 34 件の原著者推定を行っている。Benedetto らの手法を用いた実験では、著者推定に関しては高い精度が得られた。一方で、テーマ別の識別実験では、“意味論的な識別に用いうるとする主張は、疑わしいもの”と結論付けている。他に Benedetto らの手法を応用した研究としては「ヘブライ人への手紙」の著者推定実験（Sabordo ら 2005）があるが、手法の有効性が検証されていないにもかかわらず、正解著者が判明していない文書を用いた実験であり、実験の設計自体の適切性には疑問が残る。

筆者は先行研究（安形 2005）で、Benedetto らの手法の欠点を修正した圧縮改善係数による手法を提案した。そして、著者推定に定評がある従来の手法を用いた松浦ら（松浦ら 1999, 2000）の実験とほぼ同様のデータ（後述の近代日本文学者データ）に対して行った著者推定実験からの検証を行った。従来の手法と、圧縮プログラムとして ZIP を組み合わせた場合の本手法の著者推定に関する平均成功率は表 1 のようになった。この表からは、圧縮改善係数による手法の平均成功率は 97.68% であり、定評のある著者推定手法だけでなく、Benedetto らの手法よりも高い結果を出したことがわかる。

表 1　既往研究との比較

推計手法			平均成功率
松浦ら（2000）	dissim	3-gram	96.00%
	Takard の手法	2-gram	77.40%
	ダイバージェンス	1-gram	52.50%
Benedetto らの手法			90.46%
圧縮改善係数による手法（本手法）			97.68%

また、追加的な分析から ZIP の圧縮レベルにより平均成功率は変動し、先行研究と同様に圧縮率が著者推定性能に影響を与えることが明らかとなった。

圧縮プログラムを応用した類似データの同定に関して、既往研究から明らかとなった特徴は次のようにまとめられる。

1)　一般的な圧縮プログラムを利用するため導入コストが低い。
2)　テキストデータだけでなく、画像データなど、データの種類にかかわらず応用可能である。
3)　連結した全データに対して圧縮を行うため、計算量が多く大規模データには向かない。
4)　用途として、主題からの情報検索というよりは文体からの著者推定といった類似データの同定において高い性能を示す。
5)　圧縮率に関するオプションはデータの同定性能に影響を与える。圧縮率が高いほど著者推定精度が高くなる。

1. 3　本研究の目的

本研究の目的は、先行研究の成果を受け、圧縮プログラムからの著者推定手法に関する理解をさらに深めることにある。先行研究では圧縮プログラムを一つしか用いなかったが、今回は様々なアルゴリズムによる複数の圧縮プログラムを用い、前回と同様に近代日本文学データに対する著者推定実験を行った。分析の中心は、各圧縮プログラムの圧縮率と著者推定の成功率の関係である。また、著者推定における現実的な課題といえる、用いるデータの長さをより短くした場合に関しての性能分析も行った。

2　圧縮改善係数からの著者推定手法

2. 1　圧縮改善係数からの推定手法

先行研究（安形 2005）では以下のような経緯で Benedetto らの手法を

改善した圧縮改善係数からの著者推定手法を提案した。

小規模なデータに対するプレ実験から Benedetto らの提案した手法には、1）比較データだけでなく基準データの単体での圧縮されやすさが連結データのサイズに影響すること、2）連結データを連結する順序が圧縮サイズに影響すること、の二つの問題点が明らかとなった。そこで、連結データの圧縮率からデータ単体での圧縮率の影響とデータの連結順序の影響を排除する目的で、以下の数式で示される圧縮改善係数を考案した。

$$\text{圧縮改善係数} = \left(\frac{LZ_X}{L_X} + \frac{LZ_{A_i}}{L_{A_i}} \right) - \left(\frac{LZ_{X+A_i} + LZ_{A_i+X}}{L_{X+A_i}} \right) \qquad (1)$$

ここで、L はファイルサイズを示し、L_X は基準データ X のファイルサイズを、L_{X+Ai} は基準データ X と比較データ A_i を連結したファイルサイズを表している。LZ は圧縮ファイルのサイズを示しており、LZ_X は X を圧縮した場合のファイルサイズを、LZ_{X+Ai} は基準データ X を先に、比較データ A_i を後として連結した場合の圧縮ファイルサイズを、LZ_{Ai+X} は逆に連結した場合の圧縮ファイルサイズをそれぞれ表している。

式（1）は、前半が各データ単体での圧縮されやすさを、後半が連結データの圧縮されやすさを表現しており、全体として、データ単体と比較してデータを連結したことで、どの程度、圧縮率が上がったかを表している。後半部で LZ_{X+Ai} と LZ_{Ai+X} の二つを算出する理由は、圧縮プログラムのアルゴリズムと実装（バッファの大きさなど）を考慮し、二つのデータの投入順序が与える影響を排除するためである。

式（1）を基準データ、比較データのサイズが異なる場合を考慮に入れて改良したものが、次頁の式（2）になる。

圧縮改善係数は連結データの圧縮されやすさがデータ単体と比較してどの程度改善されたかを示しており、この値が高ければ高いほど、類似度が高いことを意味している。そのため、あるデータに対する類似度順の出力

$$\text{圧縮改善係数} = 2\cdot\left(\frac{LZ_X}{L_X}\cdot\frac{L_X}{L_{X+A_i}}+\frac{LZ_{A_i}}{L_{A_i}}\cdot\frac{L_{A_i}}{L_{X+A_i}}\right)-\left(\frac{LZ_{X+A_i}+LZ_{A_i+X}}{L_{X+A_i}}\right)$$

$$=2\cdot\frac{LZ_X+LZ_{A_i}}{L_{X+A_i}}-\frac{LZ_{X+A_i}+LZ_{A_i+X}}{L_{X+A_i}} \qquad (2)$$

は、基準データと各比較データのすべての組み合わせについて圧縮改善係数を算出し、値が高いものから順に比較データを並べるという手順となる。

2. 2　組み合わせた圧縮プログラム

圧縮プログラムを類似データの同定に用いる場合、圧縮サイズが得られるならばどの圧縮プログラムであっても利用可能である。

手法の原理上、圧縮率の高い圧縮プログラムを用いるほど、類似データの識別力が高くなると考えられる。そのため、どの圧縮プログラムを採用するかが重要となる。ただし、圧縮プログラムの性能は対象データの種類や特性によって変化するため、データの種類や特性に合わせた圧縮プログラムを用いる必要がある。

本研究において、圧縮改善係数による手法と組み合わせた圧縮プログラム（圧縮形式も含む）は、7-zip（PPMD）、7-zip（LZMA）、bzip2、cab、dgc durilca、ERI、gca、gzip、lha、rar、rk、sp、yz2、zip の 14 種類である。これらの圧縮プログラムは次の 4 条件のいずれかに当てはまるものを選定した結果である。

1）　テキストデータに対して高い圧縮率を示している（Maximum Compression <http://www.maximumcompression.com>）：durilca、ERI、rk など。

　原理上、圧縮性能が高くなるほど著者推定性能も高くなるはずであるが、実際のプログラムを用いた応用実験においてその傾向が現れ

るか。

2) 普及しており標準的に用いられている：bzip2、cab など。

一般的に入手しやすい圧縮プログラムを用いた場合にも現実的に応用可能な著者推定性能が得られるか。

3) 日本のプログラムであり、日本語への配慮があると考えられる：gca、lha など。

今回の実験では日本語テキストを対象とするが、日本語に特化したプログラムを用いた場合に性能に違いが見られるか。

4) 他の圧縮プログラムとは異なる特殊な圧縮アルゴリズムを用いている：sp02、yz2 など。

特殊なアルゴリズムを用いたプログラムでも著者推定が可能であるか。

ただし、7-zip に関してはアルゴリズムの異なる 2 種類のオプションを用い、zip に関しては異なる実装を 2 種類用いたため、合計で 16 通りの組み合わせとなっている。

今回の実験で数多くの圧縮プログラムを用意した理由は、外部の圧縮プログラムにデータを投入するという手法の特性上、各プログラムでの内部処理の詳細は不明であり、その分析は困難であるが、多くのプログラムによる出力結果を比較することで、各プログラムの特性を明らかにできると考えたからである。

実験に用いた圧縮プログラムの実装やオプションをまとめたものが、表 2 である。各圧縮プログラムを利用するさいには、原則として最高圧縮率となるオプションを設定している。

7-zip に関しては、数多くのアルゴリズムをオプションから選択することができるが、圧縮率が高くなる得意なデータ種が異なるため、定評がある LZMA と PPMD のアルゴリズムを、オプションで使い分けた。zip に関しては、先行研究では zip ライブラリをできるだけ圧縮率が高くなるよ

表 2　用いた圧縮プログラムの概要

ラベル	実装とバージョン	オプション	アルゴリズム
7-zip（LZMA）	7-ZIP 4.20	a -t7z -mx=9	LZ77 法に変更を加えた LZMA法
7-zip（PPMD）	7-ZIP 4.20	a -t7z -m0=ppmd: o=25 : mem=28	Dmitry Shkarin の PPMdH 法を改善した方式
bzip2	tar32.dll 2.27	-c -B9 -G	BWT 法とハフマン符号化
cab	cab32.dll 0.98	-a -o -i -ml: 21	LZ77 法、ハフマン符号化、シャノンファノ符号化
dgca	dgcac 1.08	a	gca に使ったアルゴリズムを各種ファイルに対応させたもの
durilca	DURILCA 0.4b	e -o32 -t2	PPM 法の改良版
ERI	ERI 5.1fr	a -m5	PPM 法の改良版
gca	GCA 0.9k	s	BWT 法、RangeEncoder
gzip	tar32.dll 2.27	-c -z9 -G	LZSS、LZ77 法とシャノン符号化
lha	unlha32.dll 1.96f	a -jm4	LZ77 法、ハフマン符号化
rar	rar 3.50	a -m5	非公開であるが、PPM 法などのオプションがある
rk	rk 1.04	-c -I1 -y	LZ 法と PPMZ 法
sp	sp 0.2	e -c	静的 PPM 法
yz2	yz2enc	-y	LZ77 法、RangeCoder
zip1	zip 2.31	-9	LZ77 法、ハフマン符号化
zip2	JavaSE5 のライブラリ	独自プログラム内より呼び出し	LZ77 法、ハフマン符号化

うに独自プログラムから呼び出す形で用いたが、ここでは有名な info-zip サイトからの zip 実装を最高圧縮オプションで用いたものも加えた。lha に関してはファイル配布形式としては互換性問題のため推奨されていない lha7 形式での圧縮を行った。これはより一般的な lha5 形式よりもスライド辞書が 64K バイトと大きく、圧縮率が高いためである（奥村 2003）。

用いた圧縮プログラムには sp、yz2 というあまり一般的でない圧縮プログラムが含まれる。sp は Static PPM 法を用いた圧縮プログラムであり、岡野原大輔が 2002 年度 IPA 未踏 Youth プロジェクト「単語抽出法による次世代データ圧縮法の開発」、2003 年度 IPA 未踏プロジェクト「汎用的データにおける確率的言語モデルの抽出・及びその利用」によって開発

した実験的プログラムである（Okanohara 2005）。また、yz2 は山崎敏によって開発された圧縮プログラムであり、同じ開発者によって開発された DeepFreezer という圧縮プログラムが yz1 という名前であったのを受けて、yz2 となっている。基本的なアルゴリズムは LZ77 と Range Coder の組み合わせであるが、データ列の最初の文字で LZ77 の辞書を使い分けることで、圧縮率向上を図っている点が特徴である。

3　著者推定実験の手順

3. 1　実験対象テキスト

本研究で実験集合群構築に用いたのは、先行研究（松浦ら 1999, 2000, 安形 2005）と同様に青空文庫から入手した、岡本綺堂、芥川龍之介、梶井基次郎、菊池寛、国木田独歩、水野仙子、樋口一葉、有島武郎の 8 人の近代日本文学者による 92 作品のテキストデータである。

各作品データについては、著者推定実験に用いるため、本文以外の著者、タイトル、執筆年月日などの書誌事項を除去した。また、先行研究と同様に、改行、空白は原則として一文字としているが、改行後の空白は冗長であるため除去し、半角英数記号は全角に変換している。

使用した 92 作品は、表 3 のように 72 本の小説、9 本のエッセイ、5 本の書簡形式文章、3 本の戯曲、2 本の日記、1 本の談話から構成される。

3. 2　実験集合の構築

松浦らの研究と同様の実験環境を構築するために、青空文庫からの 92 作品データを基にして固定長データの 50 の実験集合群を作成した。圧縮改善係数による手法ではデータが固定長である必要はないが、先行研究との比較を行うため、固定長のデータを作成した。

表３　実験集合に含まれる文献

著者名	タイトル	著者名	タイトル
岡本綺堂	化け銀杏	菊池寛	恩讐の彼方に
	弁天娘		勝負事
	菊人形の昔		出世
	狐と僧		忠直卿行状記
	帯取りの池		父帰る
	お照の父		藤十郎の恋
	津の国屋		若杉裁判長
	柳原堤の女		ゼラール中尉
	幽霊の観世物	国木田独歩	源おじ
芥川龍之介	あばばばば		牛肉と馬鈴薯
	アグニの神		非凡なる凡人
	秋		恋を恋する人
	あの頃の自分の事		武蔵野
	或阿呆の一生		怠惰屋の弟子入り
	或敵打の話		酒中日記
	或旧友へ送る手記		たき火
	或日の大石内蔵助		運命論者
	浅草公園 ── 或シナリオ		少年の悲哀
	一塊の土		石清虚
梶井基次郎	愛撫	水野仙子	響
	ある崖上の感情		輝ける朝
	ある心の風景		神樂阪の半襟
	泥濘		道 ── ある妻の手紙
	冬の蠅		女
	冬の日		四十餘日
	筧の話		嘘をつく日
	過古	樋口一葉	十三夜
	器楽的幻覚		にごりえ
	Ｋの昇天 ── 或はＫの溺死		大つごもり
	交尾		たけくらべ
	檸檬		うつせみ
	のんきな患者		わかれ道
	路上		ゆく雲
	桜の樹の下には	有島武郎	小さき者へ
	雪後		二つの道
	城のある町にて		片信
	蒼穹		卑怯者
	闇の絵巻		広津氏に答う
	橡の花 ── 或る私信		一房の葡萄
菊池寛	青木の出京		小作人への告別
	入れ札		水野仙子氏の作品について
	勲章を貰う話		溺
	身投げ救助業		宣言一つ
	Ｍ侯爵と写真師		想片
	無名作家の日記		私の父と母
	大島が出来る話		火事とポチ

各実験集合の作成手順は以下のとおりである。

1） 92 作品のデータ群から Mersenne Twister 法により無作為に作品を一つ選択する。
2） 作品が 30,000 文字よりも長い場合は、先頭の 30,000 文字を取り出し、実験集合に追加する。該当作品を作品プールから削除する。
3） 30,000 文字よりも少ない場合、同著者の 30,000 文字未満の作品を選択し選んだ順に連結する。30,000 文字を超えた時点で、テキストの先頭 30,000 文字を一つの実験テキストとし、実験集合に追加する。連結したすべての作品を作品プールから削除する。
4） 作品プールに作品が残っている場合、1）に戻る。作品がない場合には 5）へ進む。
5） 実験集合に作品が一つしか登録されなかった著者の場合は、著者推定が不可能となるためその著者の作品を除去する。また、著者による偏りをなくすために、一著者あたりの最大実験テキスト数を 5 とし、一つの実験集合の作成を終了する。

以上の手順で実験集合を構築したときに集合群に含まれるデータは、すべて 30,000 文字 60,000 バイトの固定長データとなる。用いるデータの長さを短くしていった場合の分析を行うため、10,000 バイトずつ短縮していったデータも作成した。

表 4 からは松浦らのデータと同様の手順で作成したにもかかわらず、特に水野仙子の値が異なっていることがわかる。無作為抽出がデータ作成手順に含まれるため、10 回データ集合を作成し、先行研究と同様の性質になるかを試行したが、そのような実験集合群は作成されなかった。

データ集合の特性に差異が見られた要因としては、

1） 青空文庫のデータに対して 1999 年時点から修正が加えられこと
2） 無作為抽出のための擬似乱数として本研究では Mersenne Twister

法（Matsumoto；Nishimura 1998）を用いていること（松浦らの研究ではどのような擬似乱数を用いたかは公開されていない）

が考えられるが、既往研究（松浦ら 1999, 2000）で公開されているデータではこれ以上の分析は行うことができない。

結果として、実験集合群の特性に若干違いは出ているが、松浦らのデータに比べ各集合に含まれる平均著者数が増加しており、著者推定の精度からはより厳しい条件となったといえる。著者「水野仙子」のテキストデータは、松浦らのデータでは半数以下の集合にのみ含まれるが、今回の集合には 3 分の 2 以上のデータに含まれている。

表４　実験集合の総計

著者名	50 集合中の合計	松浦ら（2000）
岡本綺堂	218	203
芥川龍之介	100	100
梶井基次郎	170	160
菊池寛	241	222
国木田独歩	147	129
水野仙子	88	48
樋口一葉	100	84
有島武郎	100	98
総計	1,164	1,044

3. 3　平均成功率の算出

著者推定実験の評価は、ある作家のデータと集合内の他のデータを比較し、類似度順出力を行ったとき、同じ著者の他のデータが順位 1 位に出力されれば、著者推定に成功したものとし、2 位以下に出力された場合には失敗したものとした。全推定試行数に対して、著者推定の成功数の割合を算出した。

$$\text{平均成功率} = \frac{\text{成功例数}}{\text{全推定数}}(\%) \qquad (3)$$

4　実験結果

4. 1　平均成功率

表 5 は各プログラムを用いたときの平均成功率と圧縮率を成功率が高い順に示したものである。圧縮率は元データに対する圧縮データの大きさの割合を示すため、値が小さいほど圧縮率が高いことを意味する。

表 5　平均成功率と平均圧縮率

（%）

	平均成功率	平均圧縮率
7-zip（LZMA）	99.66	36.50
rk	99.66	41.33
cab	99.66	41.85
7-zip（PPMD）	99.57	39.64
durilca	99.57	35.01
gca	99.57	38.82
lha	99.57	45.59
rar	99.57	43.30
yz2	99.57	48.33
bzip2	99.48	39.38
ERI	99.40	39.16
sp	98.37	49.05
gzip	97.68	46.07
zip	97.25	46.27
dgca	26.98	56.30

結果としては、dgca 以外のプログラムではどれも 97% を超える高い成功率が得られており、さらに、7-zip を始めとする 9 つの圧縮率の高いプログラムでは平均成功率が 99.5% 以上とほぼ全ての試行において成功していることがわかる。dgca を用いた場合の平均成功率が極端に悪い原因は dgca の圧縮率が他のプログラムに比べ低いためであると考えられる。しかし、圧縮率に関して他のプログラムと比べ dgca は 10％程度の違いしかないにもかかわらず、成功率は 70％以上の極端な差となっている。これは、

著者推定という課題に対して圧縮プログラムを応用する場合、ある一定以上の圧縮性能が必要であり、それ以下では極端に著者推定精度が低くなることを示唆している。

4. 2　データ長と平均成功率

表 6 は用いるデータの長さを短くした場合の平均成功率の変化を示したものである。松浦らの研究（松浦ら 1999, 2000）では 20,000 バイトが性能分岐点であったが、今回の実験では 20,000 バイトまでデータを短くした場合にも過半数の 8 つのプログラムで 90% を超える高い平均成功率が得られた。さらに、その半分の 10,000 バイトまで短くした場合にも多くのプログラムで 8 割を超える成功率が得られたことがわかる。このことから圧縮改善係数を用いた著者推定手法は、データ長を短くした場合にも性能劣化が少ないことがわかる。

表 6　データ長と平均成功率

（%）

	10,000	20,000	30,000	40,000	50,000	60,000
7-zip（LZMA）	83.51	92.70	96.82	98.97	99.40	99.66
cab	75.09	87.97	95.27	98.37	99.48	99.66
rk	84.97	90.89	95.62	98.11	99.23	99.66
7-zip（PPMD）	83.68	90.98	96.05	98.20	99.23	99.57
durilca	83.42	91.67	96.05	98.11	99.31	99.57
gca	81.53	87.59	87.03	95.35	98.97	99.57
lha	80.33	91.15	96.05	98.54	99.40	99.57
rar	77.84	91.92	95.45	97.77	98.11	99.57
yz2	81.79	88.66	94.83	97.63	98.97	99.57
bzip2	78.35	89.09	94.76	97.77	99.05	99.48
ERI	80.28	89.60	93.73	97.68	99.14	99.40
sp	67.10	64.52	80.93	90.98	96.99	98.37
gzip	81.44	91.15	94.16	96.13	96.65	97.68
zip	81.44	91.15	95.19	96.48	96.39	97.25
dgca	10.74	12.71	16.75	22.94	21.74	26.98

4. 3　平均成功率と圧縮率の関係

この節では、圧縮率の高低と著者推定の平均成功率の関係を検討する。データの圧縮処理では一般的に投入されるデータの長さが短くなるほど、冗長な部分が少なくなるため平均圧縮率が低くなる。そこで、今回の実験集合について圧縮プログラムに投入したデータ長と平均圧縮率の関係を表7に示す。表からはデータ長が短くなるほど平均圧縮率は低くなる傾向を読み取ることができる。

図1は表6の著者推定の平均成功率と表7の平均圧縮率を組み合わせ、データ長を変化させた場合も含め、全ての著者推定の試行に関する平均成功率と平均圧縮率のペアを散布図として表現したものである。縦軸は平均成功率を、横軸は平均圧縮率を右にいくほど圧縮率が高くなる形で示したものである。図1からは全体的に圧縮率が高くなるほど平均成功率が上がる傾向が見られた。また、相関係数を算出したところ、−0.72と強い負の相関があることが明らかとなった。

表7　圧縮プログラムの平均圧縮率

（%）

	10,000	20,000	30,000	40,000	50,000	60,000
7-zip（LZMA）	42.74	39.77	38.42	37.56	36.98	36.50
cab	48.45	45.41	43.93	43.02	42.39	41.85
rk	47.23	44.41	43.10	42.30	41.76	41.33
7-zip（PPMD）	46.20	43.02	41.58	40.70	40.12	39.64
durilca	40.87	38.17	36.88	36.05	35.49	35.01
gca	45.07	42.37	40.99	40.03	39.37	38.82
lha	51.30	48.65	47.41	46.53	46.00	45.59
rar	49.31	46.61	45.35	44.44	43.77	43.30
yz2	57.04	52.80	50.87	49.75	48.95	48.33
bzip2	45.80	42.71	41.32	40.45	39.86	39.38
ERI	45.52	42.56	41.17	40.28	39.67	39.16
sp	58.35	54.45	52.19	50.71	49.76	49.05
gzip	50.94	48.52	47.43	46.72	46.33	46.07
zip	52.22	49.15	47.78	47.01	46.57	46.27
dgca	62.39	59.22	57.99	57.25	56.71	56.30

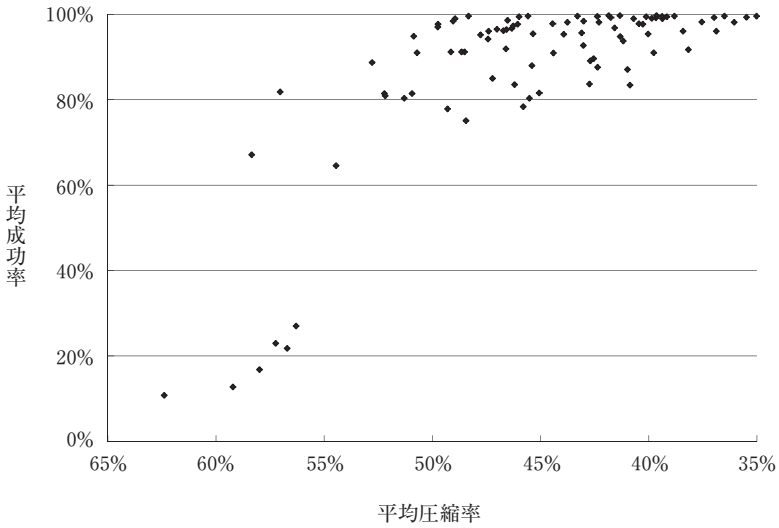

図 1　平均成功率と平均圧縮率

5　まとめ

本研究では、8 人の著者の近代日本文学データに対して様々なアルゴリズムの 16 種類の圧縮プログラムを圧縮改善係数からの著者推定手法に組み合わせた著者推定実験を行った。実験結果は以下の 3 点に集約される。

1)　圧縮性能が高い圧縮プログラムはどのようなアルゴリズムであってもほぼ 100% に近い高い著者推定の平均成功率を示す。
2)　データの長さを 20,000 バイトまで短くした場合にも圧縮性能の高いプログラムでは 9 割以上の成功率であり、さらにその半分の 10,000 バイトにしたときの性能劣化も少ない。
3)　平均圧縮率と著者推定の平均成功率には高い相関がみられた。

今回の実験では、ほぼ 100% に近い精度での著者推定を行うことができた。今後は、『時事新報』論説記事「脱亜論」を福沢諭吉が書いたか（平山 2004）など、実際に著者の真贋が問題となっている事例を取り上げ、

圧縮改善係数からの著者推定手法を応用したいと考えている。

文献

Benedetto, Dario et al.（2002）Language trees and zipping. *Physical Review Letters,* vol.88, no.4, p. 048702-1-048702-4.

Kukushkina, O.V. et al.（2000）Using letters and grammatical statistics for authorship attribution. *Problems of Transmitting of Information,* vol.37, no.2, p. 172-184（英訳版がhttp://www.philol.msu.ru/~lex/articles/grco_e.htmより入手可能）.

Matsumoto, M.; Nishimura, T.（1998）Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Transaction on Modeling and Computer Simulation,* vol.8, no.1, p. 3-30.

Okanohara, D.（2005）Okanohara, D. Partially decodable compression with static ppm. In *the Data Compression Conference 2005* poster session. Snowbird, UT, USA, March 2005.（http://www-tsujii.is.s.u-tokyo.ac.jp/~hillbig/papers/dcc2005.pdfより入手可能）

Tankard, J.（1986）The Literary detective. *BYTE*, vol.11, no.2, p. 231-238.

Sabordo, M. et al.（2005）Who wrote the Letter to the Hebrews? - Data mining for detection of text authorship. *Smart structures, devices, and systems II* : 13-15 December 2004, Sydney, Australia / Said F. Al-Sarawi(ed.), p. 513-524.

Zheng, Rong et al.（2006）A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, p. 378-393.

安形輝（2005）圧縮プログラムを応用した著者推定. *Library and Information Science*, no.54, 2005, p. 1-18.

奥村晴彦（2006）データ圧縮.『数学セミナー』, vol.45, no.10, p. 28-34.

奥村晴彦監修（2003）『LHAとZIP――圧縮アルゴリズム×プログラミング入門』. ソフトバンク パブリッシング, 258p.

佐藤進也ほか（2002）文字列出現頻度比較による情報源間の類似性判定.『情報処理学会研究報告』, vol.2002, no.028（FI-066）, p. 119-126.

細江光（1988）谷崎の作品ではなかった偽作「誘惑女神」をめぐって.『国文学 解釈と教材の研究』, vol.33, no.8, p. 134-137.

フリードマン, リチャード・エリオット（著）, 松本英昭（訳）（1989）『旧約聖書を推理する――本当は誰が書いたか』. 海青社, 355p.（原書の再版が1997年に刊行さ

れている）.
松浦司; 金田康正（1999）近代日本文学者8人による文章における文字N-gram分布を手がかりとする著者推定.『情報処理学会研究報告』, vol.99, no.95（NL-134）, p. 31–38.
松浦司; 金田康正（2000）n-gramの分布を利用した近代日本語文の著者推定.『計量国語学』, vol.22, no.6, p. 225–238.
森直久（2001）ある刑事事件の供述資料における作成者同一性の心理学的検討.『札幌学院大学人文学会紀要』, vol.69, p. 13–36.
村上征勝（2002）著者を探る古文書の計量分析.『電子情報通信学会誌』, vol.85, no.3, p. 158–161.
村上征勝（1994）『真贋の科学 ── 計量文献学入門』. 朝倉書店, 154p.
内山和也（2002）スタイルの計量に関する覚え書き ── 文体論の視点から.『計量国語学』, vol.23, no.7, p. 347–352.
平山洋（2004）『福沢諭吉の真実』. 文藝春秋 , 244p.（文春新書 394）